A Next Generation Information System for Earth Science Data

Rahul Ramachandran
Helen T. Conover
Sara J. Graves
Ken Keiser
Charles Pearson
John Rushing

Information Technology and Systems Center
University of Alabama in Huntsville

ABSTRACT

Many scientific data archives today are highly inflexible collections of static data objects that can be retrieved and distributed, but not easily manipulated to suit the users' requirements. Scientists, as well as their analysis and visualization tools, are significantly restricted by the limitations of data archive centers. These limitations include inability of the data centers to search archives based on the data content, inaccurate and ineffective geographic searches on swath data, inability to subset data products, lack of custom data products, lack of choices for gridding schemes and limitation of format choices. The Next Generation Information System being developed under the auspices of NASA's Passive Microwave Earth Science Information Partner (PM-ESIP) will address all the above limitations. This innovative system is being built around the flexible and extensible Algorithm Development and Mining System (ADaM) developed by the Information Technology and Systems Center at the University of Alabama in Huntsville. ADaM provides an extensible framework for data mining and other types of processing including subsetting, gridding and mapping. The main emphasis for the design of the new PM-ESIP information system is to allow the user flexibility and ease in accessing and utilizing data. The PM-ESIP's next generation information system will be a scaleable distributed processing system that can grow as the need of the user community for on-demand processing increases. This paper will discuss the design of this system along with some of the technology being used to build it.

Keywords: data mining, custom order processing, on-demand processing, subsetting, ESIP

1. INTRODUCTION

There is a wide variety of potential users of Earth Observing System (EOS) data, ranging from the most sophisticated and technologically savvy scientist to the engineer or policy maker interested in using EOS data to answer questions in agricultural productivity, forest resource management, watershed management, or utility needs. Many scientific data archives today are highly inflexible collections of static data objects that can be retrieved and distributed, but not easily manipulated to suit the users' requirements. Scientists, as well as their analysis and visualization tools, are significantly restricted by the limitations of data archive centers. These limitations include inability of the data centers to search archives based on the data content, inaccurate and ineffective geographic searches on swath data, inability to subset data products, lack of custom data products, lack of choices for gridding schemes and limitation of format choices.

The Next Generation Information System being developed under the auspices of NASA's Passive Microwave Earth Science Information Partner (PM-ESIP) will address all the above limitations. This innovative system is being built around the flexible and extensible Algorithm Development and Mining System (ADaM) developed by the Information Technology and Systems Center at the University of Alabama in Huntsville. Data mining can unearth nuggets of valuable information in mountains of scientific data, thus allowing researchers to explore, analyze, visualize, and summarize the data contents. ADaM provides the tools necessary to perform a variety of types of data mining, including using well-defined scientific algorithms to search for particular events; using pattern recognition techniques to develop new algorithms for known events; and searching through data for unknown relationships and transient events. In addition to data mining, ADaM provides an extensible framework for other types of processing including subsetting, gridding and mapping.

The main emphasis for the design of the new PM-ESIP information system is to allow the user flexibility and ease in accessing and utilizing data. The system will offer a number of different data search perspectives in addition to a standard instrument/parameter-based search to assist the user in determining the particular data set of interest. It will also provide a suite of innovative tools for custom order processing. For some data sets, the PM-ESIP system will contain a variety of scientific algorithms for on-demand production of higher-level products from instrument data. This system will emphasize subsetting capabilities, including spatial and temporal subsetting, subsetting based on channel or parameter, on coincidence with other data sets, or on data values. Mapping and gridding services will be available on this system to provide the data in a format and resolution that fits the user's need. The PM-ESIP's next generation information system will be a scaleable distributed processing system that can grow as the need of the user community for on-demand processing increases. This paper will discuss the design of this system along with some of the technology being used to build it.

2. FEATURES OF THE NEXT GENERATION INFORMATION SYSTEM

The Next Generation Information System (NGIS) will provide a suite of innovative order processing tools to the end user. It will offer a number of different data search perspectives in addition to a standard instrument/parameter based search to assist the user in determining the particular data set of interest. Some of the key features that will make the system user friendly are:

2.1 Searching Archives Based on Data Content

Queries within most current earth science archives are restricted to searching predefined parameters or metadata, which point to a particular set of data files. Often these focus on general data set characteristics such as satellite ID, sensor type, area and time sampled or geophysical parameter measured rather than the data value. In-order to extract information out of data sets, the NGIS will allow the user to search the content of the files. This will be particularly useful for analysis such as extricating scientific phenomena and abnormalities in data set.

2.2 Subsetting Data Products

Many data management systems treat data files stored within the archive as the smallest deliverable unit of data. File size may be optimized for archive (typically 50 MB or more) rather than for data analysis. Delivering such large files to the end user when small subsets of the data are desired places unnecessary burdens on the network, on local storage capacities and on the user who must then sift through the data to access the desired data subset. Subsetting data products will be one of the salient features of NGIS and will provide substantial cost savings to both the data center and the science user.

2.3 Custom Data Products

With most current earth science data products, decisions regarding algorithms, grid type and resolution, and temporal slicing or averaging have been made by the data producer or by a data committee. However, the production of ready made higher level data products has come at a cost of limited data products, algorithms that may be inappropriate for the desired application or data products that have insufficient temporal or spatial resolution to be useful for many regional applications. Typically, for pre-calculated data sets, the data production site must balance the desires for higher resolution and greater number of products against available CPU processing and disk storage capacities. The PM-ESIP NGIS is moving away from this static approach to a more dynamic one. Product on demand and custom order processing generate the product when it is wanted with user specified processing and packaging. These capabilities lay the foundations for the NGIS.

2.4 User-Selectable Choices for Gridding Schemes

The lack of common mapping and gridding schemes can result in significant difficulties in the inter-use of multiple data products. Although most data sets can be transformed between various map projections, and can be resampled to account for resolution differences, these transforms can result in significant distortion and degradation of data, perhaps to the point where the data values are no longer scientifically valid. Ideally, if a given gridding scheme is required or desired for an application, one would want to transform and resample the data from the native sensor space to the desired grid. The NGIS will have this capability of providing gridding options based on full-resolution data to the end user.

2.5 Data Format Choices

The use of standard data formats such as Hierarchical Data Format (HDF) and HDF modified for use in EOS (HDF-EOS) provides benefits to many users and can greatly simplify tool development in support of EOS data sets. However, the lack of user specified data formats can also place extra burden on those users whose tools may not support HDF-EOS or for whose purpose a standard data format might be overkill. Choices of different data format are another key characteristic that NGIS will provide. This removes the onus of data conversion from the end user.

2.6 Distributed Data Processing

The era of data crunching on large single mainframes is over. More powerful desktop computers together with the improvements in networking technology, provide the means for resource sharing in a distributed setting. The NGIS will be able to maximize the usage of resources available to it. One of the ways of achieving this, especially for CPU intensive operations, is to have the ability to process jobs on machines at different locations.

3. DESIGN OF THE PM-ESIP PROTOTYPE

Figure 1 depicts the conceptual design of the PM-ESIP prototype. The user is provided various choices for searching the data archive including search by parameter values or geo-temporal ranges and an iterative coincidence search system. The data archive contains data sets such as Tropical Rainfall Measuring Mission Microwave Imager (TMI), Advanced Microwave Sounding Unit (AMSU), Special Sensor Microwave/Imager (SSM/I), etc. Once data set selection has been made, the data set is retrieved from the archive. The user is then presented with multiple processing options such as gridding, subsetting, format conversion, etc., for customizing this data set.

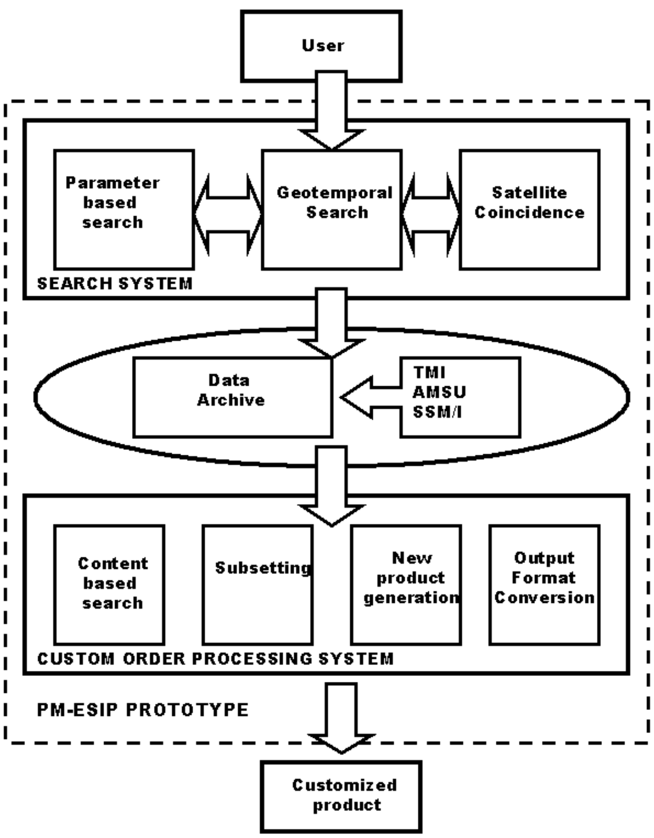

Figure 1: Conceptual design of the PM-ESIP Prototype

Figure 2 represents the system architecture of the PM-ESIP prototype. The processing engine of this prototype is built upon the Algorithm Development and Mining system (ADaM) developed by the Information Technology and System Center at the University of Alabama in Huntsville. Details on this processing system will be provided in the next section. The end user's World Wide Web browser serves as a gateway to this prototype. The three-tier architecture has an application layer with built-in intelligence derived from the metadata stored in a database. This application layer is comprised of two components. One of the components, the Database Communicator queries the database for metadata and provides the users different options based on their selection. This metadata contains the information about all the data sets and the associated operations that apply to them. The second component, the Order Translator, converts the user selections and specifications into mining and custom processing jobs. The jobs are then passed by the Order Translator component to servers (processing engine) residing at different locations. This design allows the user to take advantage of the capabilities of the ADaM system and also permits distributed processing across a network.

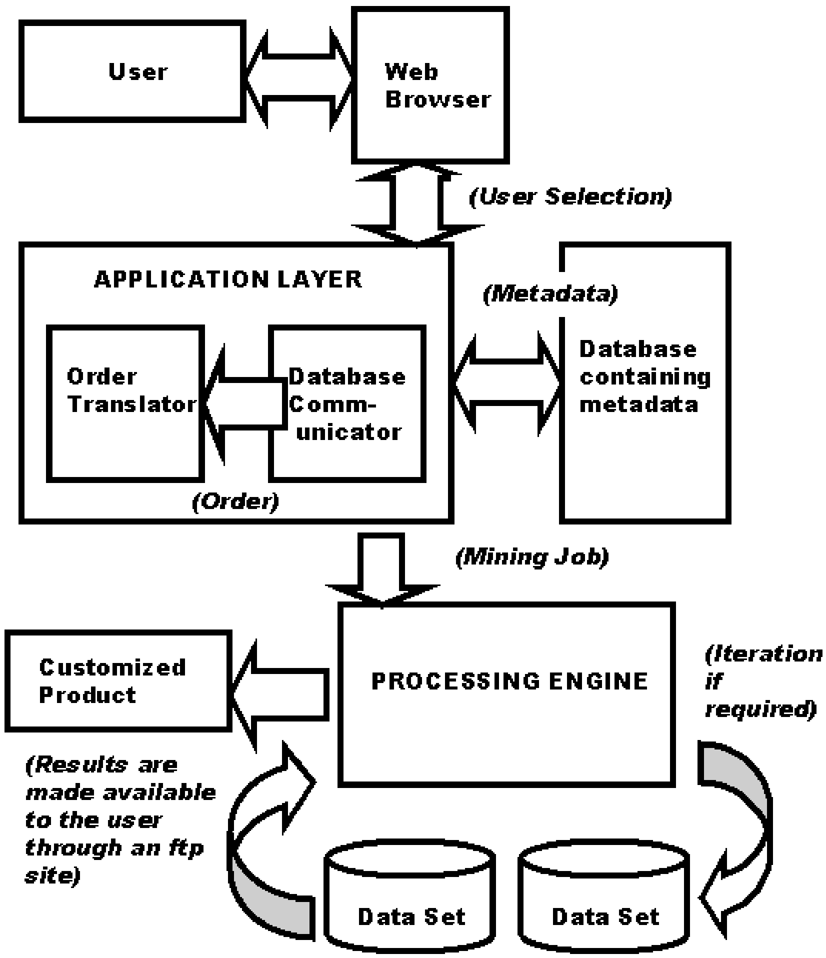

Figure 2: PM-ESIP Prototype system architecture

4. ADAM: ALGORITHM DEVELOPMENT AND MINING SYSTEM

Data Mining is a technique that allows a user to scan very large databases to retrieve the high level information of the greatest interest. These mining discoveries could vary from a simple keyword pattern search to subtle trends and clusterings. ADaM was developed in response to the need to mine large scientific data sets for geophysical phenomena detection and feature extraction. It provides a variety of processing tools, which allow easy integration of spatial and temporal variables of earth science data sets. The system was initially developed to provide searching on data values as well as on metadata, and to catalog the information discovered to enhance the existing metadata. Algorithms that detect a variety of geophysical phenomena were added to the system to address specific needs of the earth science community. Content-based metadata resulting from mining operations can be stored to allow searches by researchers via web-based interfaces. Results from searches on this enhanced metadata can point researchers to the actual data sets (Hinke et al, 1997). Many of ADaM's data operators are useful for general processing as well as data mining. The system's flexible, modular design makes it an excellent framework for a variety of specialized applications, such as the PM-ESIP NGIS.

4.1 Design Features

ITSC developers applied the latest object oriented software design concepts while developing the ADaM system. These design features allow ADaM to be utilized effectively as a backend for the NGIS. The key features of the design of the system are:

- **Portability -** In order to realize a high degree of platform independence, the system was written using widely available, standard tools such as the C++ programming language and ANSI C libraries. This has allowed the Center to develop the same system for Windows and UNIX operating systems.

- **Network Accessibility -** The client-server architecture of the system allows ease in network accessibility. This feature of the system allows it to be used as an application at a data archiving center or on a user's desktop workstation.

- **Extensibility -** The ADaM system consists of three basic types of modules: input filters (readers for different data formats), processing modules (general-purpose algorithms and user-defined algorithms) and output filters (writers for different data formats). Since the number of data sets and algorithms for analysis keep increasing and changing, the system was designed to be extensible by the use of plug in modules. New modules can be added to the system easily. Thus, the system is designed to evolve along with the demand.

4.2 Processing Architecture

The ADaM system architecture utilizes a data pipeline approach. Mining is broken down into a series of steps with results from each step passed to the next one in line. Figure 3 illustrates both ADaM's data processing stream, as well as the three basic types of modules: input, processing, and output. The use of data input filters, specialized for a variety of data types, has been instrumental in simplifying the development of the processing and output operations. The selected input filter translates the data into a common internal structure so that the processing operations can all be written to a single data representation. This allows the addition of new operations to the system without having to address input data format problems. Similarly, the addition of a new input filter provides access to the entire suite of processing operations for the data type in question. The mining system currently allows over 80 different operations to be performed on the input data stream. These operations vary from specialized atmospheric science data set specific algorithms to generalized image processing techniques. The last step in the mining process is the selection of the output format. Since the input data has been converted to ADaM's internal format, the output modules allow the user the option to select either the input format or a different format for the final data product. In the same manner as the input modules, the output filters effectively insulate the processing operations from having to support all the possible output formats. Details about the ADaM System can be found in Keiser et al (1999).

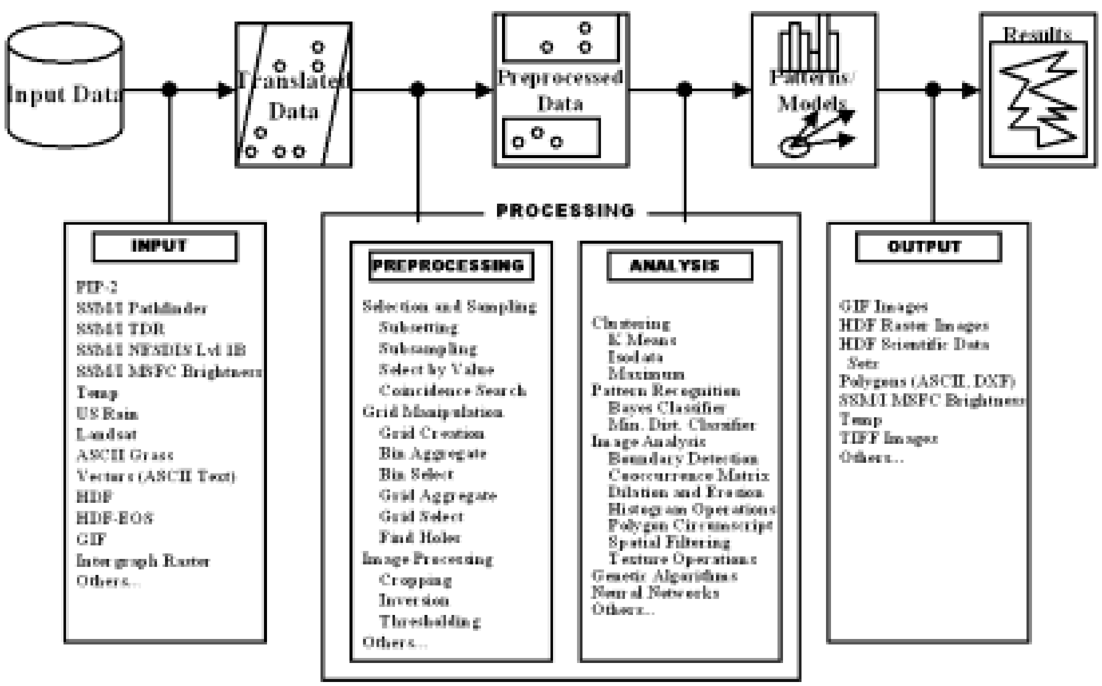

Figure 3: Schematic diagram depicting the processing architecture of the mining system

4.3 Mining Occurrence

Mining or other processing using ADaM can occur at various times in a data production workflow. The time chosen depends upon the problem being addressed. The possible times include:

- **Real Time -** In some cases data mining must occur in real time in order for the data to be of any use in the application. This is a common requirement for applications that predict future events such as severe weather. For example, the PM-ESIP is using ADaM for tropical storm detection and maximum wind speed estimation in near real time AMSU-A data.

- **On Ingest -** Certain types of data mining are always performed on data to make it useful. Such mining or processing could include navigation, quality control, generation of enhanced metadata, etc. Such processing is done most efficiently on ingest, before the data is archived.

- **On Demand -** ADaM can provide content-based search or custom order processing for the end user, as needed. For example, PM-ESIP users may want gridding, subsetting, subsampling and other operations to be performed to suit their specific needs.

5. CAPABILITIES OF THE PM-ESIP PROTOTYPE

The PM-ESIP prototype will use the ADaM system to provide a suite of innovative order processing tools to the end user. This section discusses how each of the key features of the NGIS, enumerated above, is addressed by the PM-ESIP's use of ADaM:

5.1 Searching Archives Based on Data Content

For any data product in the archive and for predefined virtual products that can be generated on demand, the PM-ESIP prototype supports two types of search, existence and content. The existence search will indicate whether a particular data product is available coincident with a specified temporal range and/or spatial area, while the content based search will allow the user to search or mine for particular data values or geophysical phenomena within a specified data product. The content based search feature will be provided by the existing data mining system.

Content based search can be very resource and time intensive due to the need to retrieve data from tertiary storage, decompress it, and scan the individual data points for values or phenomena of interest. The PM-ESIP prototype utilizes a number of search strategies to mitigate this high resource intensity. These strategies include prescreening, or performing the content-based search during initial data ingest; controlled processing of archived products on tertiary storage or of virtual products that must be created prior to being scanned; and multiprocessing. The multiprocessing will be done using a number of PC or UNIX-class machines receiving data to search over a high-speed local area network. Some content-based searches, such as for the presence of precipitation in a particular space time window can be performed on the browse images rather than on the full resolution data products. Because the browse images will generally be kept online, and because they will be much smaller than the full resolution data products, this content based search will be more efficient.

5.2 Subsetting Data Products

Subsetting data products should provide substantial cost savings to both the data center and the science user. This prototype provides subsettting for map-projected and native swath products based on spatial or temporal criteria, data parameter, or coincidence searches based on geographic features. Subsetting is achieved by either utilizing the subset routines built into the readers for certain data sets or by invoking ADaM's subsetting operation. It also takes subsetting one or two steps further by allowing the user to define the grid or map projection of the resulting data file. Temporal subsetting and gridding will also be supported, permitting seasonal, monthly, pentad or hourly averages, or user-defined time periods. Figure 4 is a screen shot of the interface that allows the user to perform customized subsetting based on geo-temporal parameters.

5.3 Custom Data Products

The ability of the NGIS to constantly adapt is the direct benefit of employing ADaM as the processing engine of the information system. This system is very innovative and evolutionary. As stated earlier, addition of input, processing or output modules to ADaM is relatively easy and does not require rebuilding the entire system. Thus, NGIS can easily add new data sets and different operations to customize data products based on user need and specifications.

5.4 User-Selectable Choices for Gridding Schemes

This prototype provides gridded data on demand. This feature is again based on the gridding operations available in the ADaM. The user is now able to request that any data product, stored or virtual, be mapped, gridded, and subsetted to his or her specifications. Figure 5 is a screen shot of the interface where the user can select gridding options.

5.5 Data Format Choices

The clear separation of the input, processing and output modules in ADaM allows the reader to read data sets in different formats, apply a variety of analysis algorithms to the data, and finally convert it to either its native format or any other format desired by the user. Thus, this prototype information system supports a variety of distribution data formats (e.g., GIF, TIFF, GeoTIFF, JPEG, raw ASCII, binary files or others) at the data server. The screen shown in Figure 5 also allows the user to select output formats.

5.6 Distributed Data Processing

ADaM was designed as a client-server architecture to provide network accessibility. The central processing engine and client applications communicate through a mining daemon component. This feature of ADaM allows it to be utilized for distributed processing (Ramachandran, 1999). The central mining engine can be on a local or remote server. The clients can connect to a server from different locations for processing their jobs. A client can also access multiple servers on different machines. This capability is utilized in NGIS for distributed data processing.

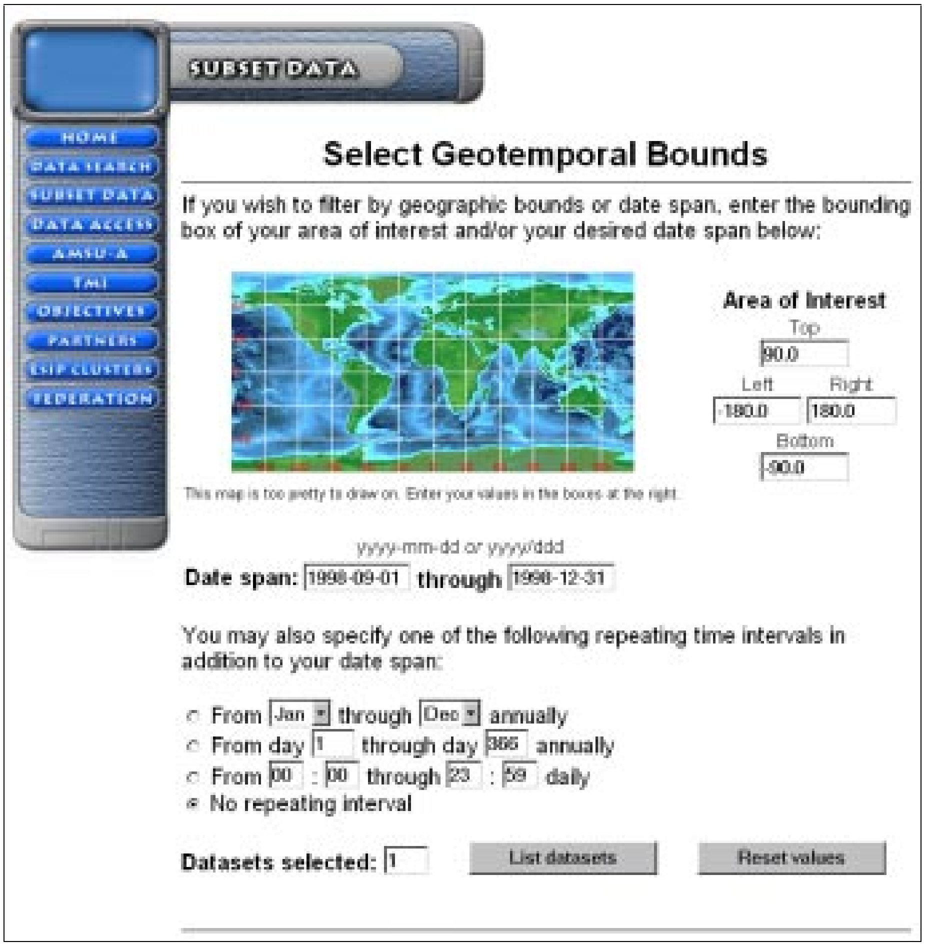

Figure 4: Screenshot of the interface providing subsetting options

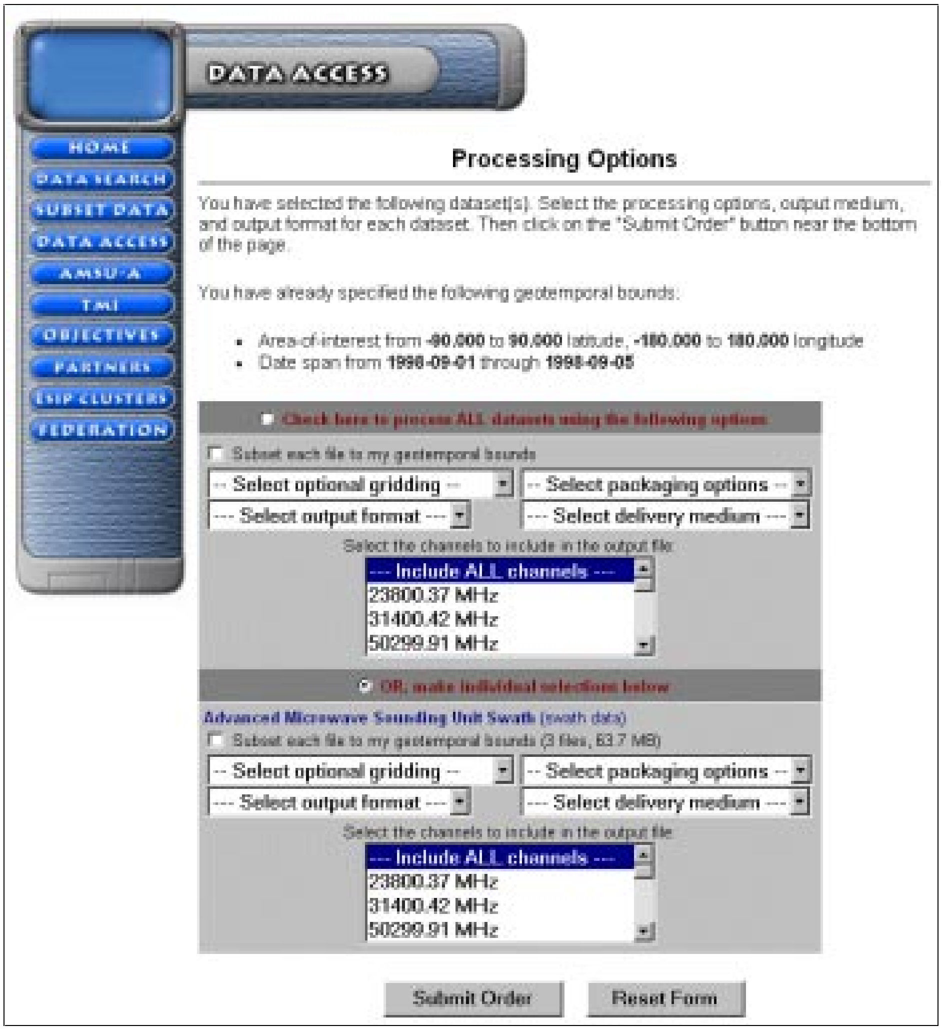

Figure 5: Screenshot of the interface providing processing (gridding and output format) options

6. PHENOMENON DETECTION IN THE PM-ESIP

In addition to custom order processing, another aspect of the PM-ESIP's Next Generation Information System is the Tropical Cyclone Windspeed Indicator, also based on AdaM. The strong surface winds in tropical cyclones are directly related to the warm middle- and upper-temperatures which exist around the cyclone center. Using a method developed by the PM-ESIP science team, maximum sustained winds are estimated using gradients in temperature data from the Advanced Microwave Sounding Unit (AMSU-A) instrument flying on the NOAA-15 satellite. The wind speed data is then analyzed to detect tropical cyclones. The method has been calibrated using aircraft reconnaissance measurements in tropical depressions, tropical storms, and hurricanes from the 1998 Atlantic hurricane season.

As AMSU-A data is ingested into the PM-ESIP, it is processed by AdaM using a combination of general-purpose image analysis modules and special purpose modules developed specifically for this problem. The resulting images are subsetted to the areas around each storm for display on the web. Figure 6 depicts the data flow in the tropical cyclone detection process. The near-real-time storm information can be viewed at http://pm-esip.msfc.nasa.gov/cyclone/.

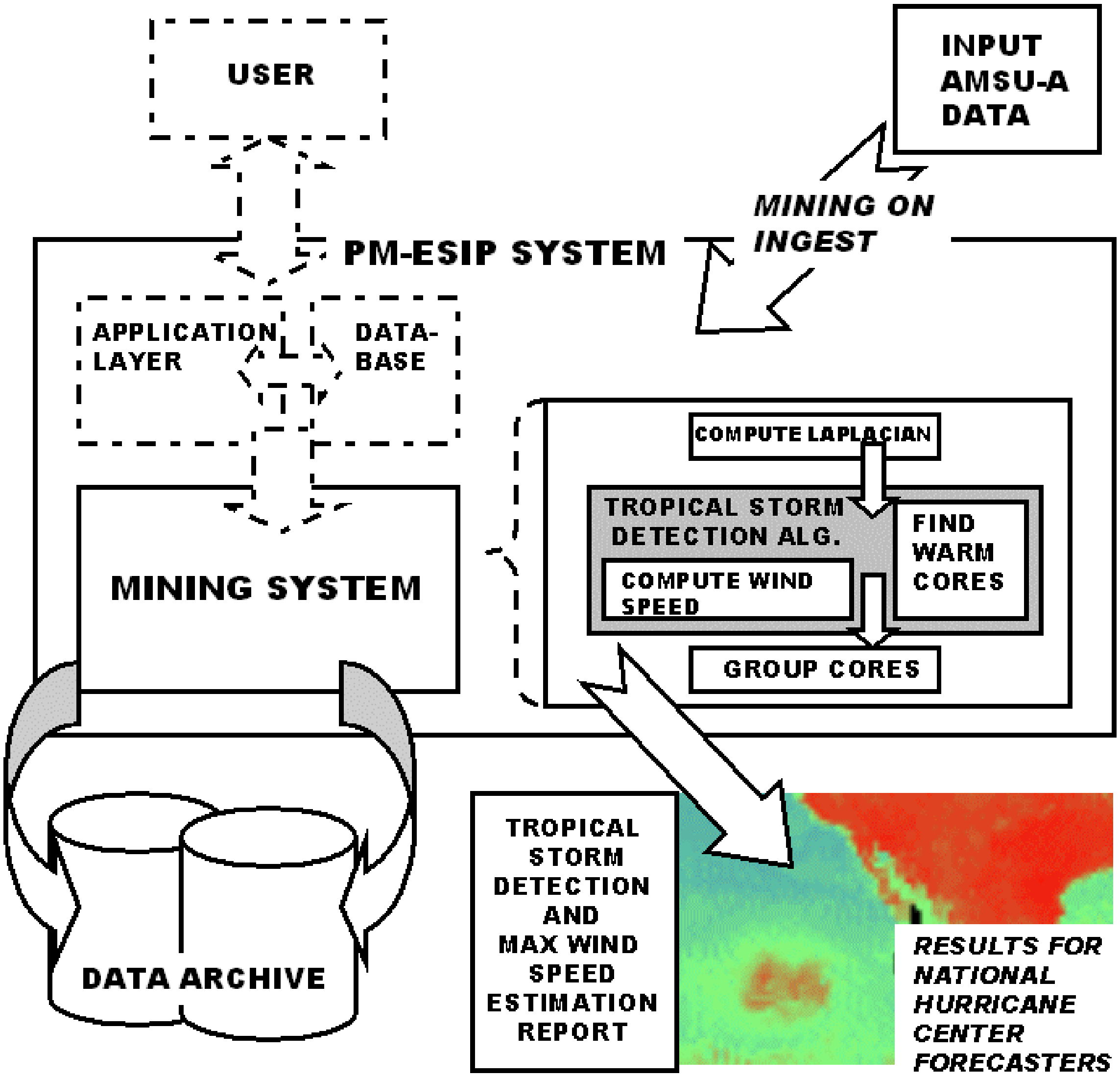

Figure 6: Tropical cyclone detection in the PM-ESIP.

7. SUMMARY

The NGIS brings about a shift in paradigm for data distribution centers. Instead of being static, inflexible and producer biased data providers, the emphasis for NGIS is now on the user. The NGIS is user friendly to allow maximum access and utilization of its data products and services by end users. The features of the NGIS are: ability to allow geographic searches on swath data, search data archives based on data content, subset data, customize the data products, provide gridding schemes and output format choices.

The PM-ESIP prototype is an example of how the innovative use of technology is used to usher in the new paradigm. This prototype utilizes the ADaM system to its full advantage. It uses all the features that make ADaM unique: ADaM's data format independence, its gridding and subsetting operations, its ability to add new operations, readers and writers for different data sets and needs and finally, its basic ability to provide mining (content-based searches) within data files itself.

ACKNOWLEDGEMENTS

This study was conducted under National Aeronautics and Space Administration grant NCC 8-141 and NAGW-4259.

REFERENCES

Hinke, T., J. Rushing, S. Kansal, S. Graves, H. Ranganath and E. Criswell, 1997. Eureka Phenomena Discovery and Phenomena Mining System. *13th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology*, February 1997.

Hinke, T., J. Rushing, S. Kansal, S. Graves, and H. Ranganath, 1997. For Scientific Data Discovery: Why Can't the Archive be More Like the Web, *Proc. 9th Int. Conf. On Scientific Database Management*, Olympia, WA, Aug. 11-13, 1997.

Hinke, T., J. Rushing, H. Ranganath and S. Graves, 1997: Target-Independent Mining for Scientific Data: Capturing Transients and Trends for Phenomena Mining, *Proc. Third Int. Conf. On Data Mining (KDD-97)*, Newport Beach, CA, Aug. 14-17, 1997.

Keiser, K., J. Rushing, H. Conover and S. Graves, 1999. Data Mining System Toolkit for Earth Science Data. *Earth Observation and Geo-Spatial Web and Internet Workshop (EOGEO)-1999*, Washington, Feb 9-11.

Ramachandran, R., H. Conover, S. Graves, K. Keiser and J. Rushing, 1999. The Role of Data Mining in Earth Science Data Interoperability, *ASPRS Annual Conference, Conference on Remote Sensing Education (CORSE), Education for the Next Millennium*, Portland, Oregon, May 17-22, 1999